Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules

Harmen J. Bussemaker, Barrett C. Foat, and Lucas D. Ward

Department of Biological Sciences, Columbia University, New York, New York 10027; email: hjb2004@columbia.edu

Annu. Rev. Biophys. Biomol. Struct. 2007. 36:329–47

First published online as a Review in Advance on February 20, 2007

The *Annual Review of Biophysics and Biomolecular Structure* is online at biophys.annualreviews.org

This article's doi:
10.1146/annurev.biophys.36.040306.132725

1056-8700/07/0609-0329$20.00

Key Words

transcriptional and posttranscriptional regulation, *cis*-regulatory logic, quantitative modeling, sequence specificity, transcription factor activity

Abstract

Various algorithms are available for predicting mRNA expression and modeling gene regulatory processes. They differ in whether they rely on the existence of modules of coregulated genes or build a model that applies to all genes, whether they represent regulatory activities as hidden variables or as mRNA levels, and whether they implicitly or explicitly model the complex *cis*-regulatory logic of multiple interacting transcription factors binding the same DNA. The fact that functional genomics data of different types reflect the same molecular processes provides a natural strategy for integrative computational analysis. One promising avenue toward an accurate and comprehensive model of gene regulation combines biophysical modeling of the interactions among proteins, DNA, and RNA with the use of large-scale functional genomics data to estimate regulatory network connectivity and activity parameters. As the ability of these models to represent complex *cis*-regulatory logic increases, the need for approaches based on cross-species conservation may diminish.

Contents

INTRODUCTION

The nucleus of a cell may be viewed as a molecular computer that processes dynamic regulatory inputs according to a program defined by the static genome sequence; the outputs are the expression levels of all genes, which together define the phenotype of the cell. The computation is performed in parallel everywhere along the chromosomes. DNA carries the genetic information and, together with a complex mixture of protein, RNA, and other molecules, self-organizes into a three-dimensional chromatin structure that carefully orchestrates gene expression. The chromatin responds dynamically to changes in the cell's environment that are relayed to the nucleus by a variety of signaling pathways. Chromatin structure is also affected by genetic variation among individuals in coding and noncoding sequence.

Hidden variable: a quantity that is not measured directly but whose effects can be inferred from measured quantities

The complete genome sequences for a variety of organisms have been determined. In addition, high-throughput functional genomics technologies make it possible to probe the state of the nucleus in different ways. DNA microarrays are a particularly useful tool for measuring not only the mRNA expression level for all genes, but also the in vivo occupancy of the DNA by hundreds of different DNA-binding and other chromatin-associated proteins. Because all these data reflect the same underlying molecular processes, much will be gained by modeling them in an integrated fashion (**Figure 1**).

In recent years, significant progress has been made toward the construction of a biophysically motivated in silico model that can quantitatively predict the response of the nucleus to a variety of genetic and environmental perturbations. Such a model will deepen our understanding of organismal development and cellular physiology. It will also be of value for understanding disease processes, designing novel drugs and strategies for personalized medicine, and de novo engineering of gene regulatory networks in microorganisms.

Many molecular players involved in the regulatory processes inside the nucleus are known. However, quantitative information about their interactions with the DNA and each other is far from comprehensive. Because these parameters can be observed only indirectly through the available high-throughput functional genomics data, quantitative modeling is required. First, the binding energies associated with specific protein-DNA and protein-protein interactions (the strength of the arrows in the network) must be determined. Second, the activities of hundreds of regulatory proteins (the nodes in the network), many of which are dependent on the cellular state, are hidden variables that must be estimated from the expression levels of the genes they control.

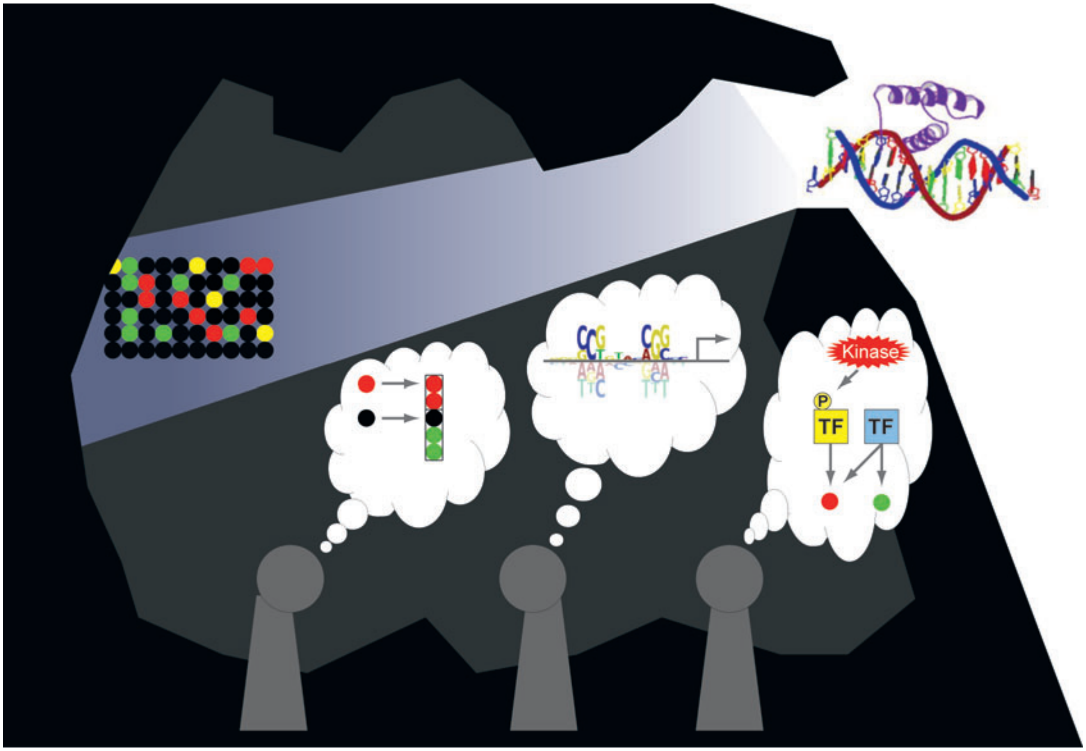

Figure 1

Gene expression modelers as prisoners in Plato's Cave. They can observe the biomolecular processes in the cell only indirectly, via high-throughput experiments; therefore they each build their own imperfect representation of reality.

Available High-Throughput Data

Over the past decade, genome-scale technologies have matured from novelties to ubiquitous components of molecular biology research. Whole-genome sequencing projects for all popular model organisms have been completed, and a variety of whole-genome functional assays based on microarray technology have been developed. The first use of microarrays focused on expression profiling of all genes in the yeast genome in various conditions and genetic backgrounds (21, 40). Subsequently, microarray techniques were adopted to determine the in vivo occupancy profile by transcription factors (TFs) along the genome. These assays are based either on chromatin immunoprecipitation microarray experiments (ChIP-chip) (42, 70) or on DNA adenine methyltransferase identification (DamID) (92). Recently, investigators used ChIP-chip to map the RNA polymerase complex (48), the position of nucleosomes (97), and the acetylation and methylation state of histone tails (68). High-throughput methods for probing DNA-protein interactions in vitro are also available (55, 62).

Although the majority of genomic studies have focused on transcriptional regulation, posttranscriptional processes related to mRNA can also be probed using microarrays. The relative abundances of alternatively spliced transcripts have been measured extensively (11). Localization of mRNAs to various parts of the cell can also be assessed using microarrays (34). Researchers have studied the regulation of translation by isolating transcripts associated with different numbers of

TF: transcription factor or *trans*-factor (if referring to both DNA- and RNA-binding proteins)

ChIP-chip: chromatin immunoprecipitation microarray experiment

DamID: DNA adenine methyltransferase identification ribosomes and measuring the relative abundances of mRNAs on a microarray (59, 98). Genome-scale methods have also been applied to the regulation of mRNA stability (59). Most genomic approaches to studying mRNA decay arrest transcription (39, 50) and then perform time courses to infer half-lives for every transcript. Using a run-on method, investigators can compare steady-state mRNA abundances with transcription rates to infer mRNA stabilities (27). Finally, methods analogous to ChIP-chip can be used to identify the target mRNAs for particular RNA-binding proteins (89).

TF activity: concentration of the actively regulating form of a *trans*-factor, which depends on posttranslational modifications, interactions, and subcellular localization

MODELING TRANSCRIPTION FACTOR–DNA INTERACTION

TFs are central players in the regulation of gene expression. By acting as adaptor proteins between specific sites on the genome and various enzymatic complexes such as RNA polymerase, histone-modifiers, and chromatin-remodelers, they orchestrate genome-wide expression (see **Figure 2**). The modular organization of TFs, comprising a DNA-binding domain (DBD) and other domains that mediate interaction with cofactors and communication with the signaling machinery of the cell, has allowed evolution to optimize gene-specificity and condition-specificity independently. The same modularity can guide the modeling of gene regulatory processes. Before discussing how the condition-specific modulation of TF activity can be modeled, we outline approaches for quantifying the sequence specificity of the DBD of a TF, which determines, in part, the extent to which a TF controls the expression of a gene.

As a first step toward modeling the genome-wide in vivo occupancy pattern of a TF in terms of the interactions with its binding partners and their concentrations, it is important to have an accurate quantitative understanding of the interaction between the purified protein and naked DNA. Most modeling studies treat a TF-target relationship as something that either does or does not exist, and even when quantitative information about the TF-DNA interaction is available from experiments or computational analysis, a threshold is usually imposed to enforce a binary answer. From a biophysical point of view, however, a quantitative mathematical description of TF-DNA interaction is more natural. In thermodynamic equilibrium, a simple nonlinear relationship exists between the free-energy gain associated with

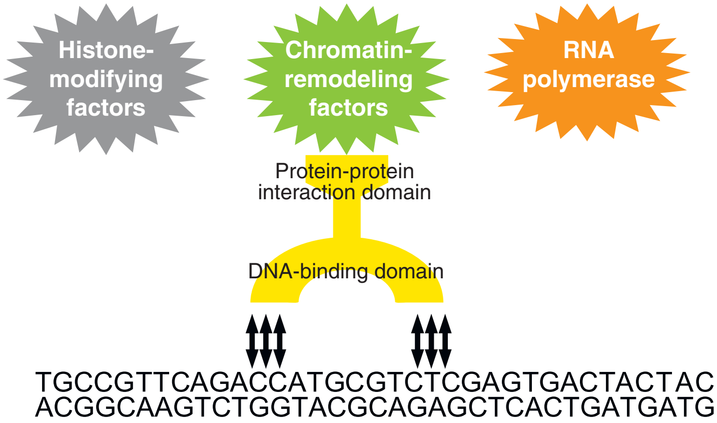
Histone-modifying factors
Chromatin-remodeling factors
RNA polymerase
Protein-protein interaction domain
DNA-binding domain
TGCCGTTCAGACCATGCGTCTCGAGTGACTACTAC
ACGGCAAGTCTGGTACGCAGAGCTCACTGATGATG

Figure 2

Transcription factors (TFs) serve as adaptors between specific genomic loci and the various enzymatic complexes that can either promote or repress gene expression. Complex interactions with other TFs and chromatin-associated proteins make it challenging to model this process accurately for a living cell.

TF-DNA binding, the concentration of the TF, and the occupancy (or fractional saturation) of the DNA binding site by the TF (see Sidebar). Every site in the genome, in principle, is targeted by a given TF, but there is a wide range of affinities (and therefore TF occupancies) associated with the variation in local DNA sequence along the genome. Indeed, a recent study showed evidence of conservation across species of the quantitative affinity level, even for suboptimal TF binding sites that would have fallen below the threshold in a discrete approach to modeling regulatory network topology (86).

Do Weight Matrices Represent DNA or Protein?

TFs bind to different DNA sequences with different efficiencies (82). A classic paper by Berg & Von Hippel (8) lays out a theoretical framework for inferring a model for sequence specificity from a collection of experimentally determined TF binding sites. It is important to recognize that this framework consists of two distinct layers: the first pertaining to the thermodynamics of TF-DNA interaction and the second related to evolutionary selection. The central assumptions are that natural selection has given rise to a certain level of sequence specificity for each TF and that sequences that give rise to the same physical binding affinity are equally likely to be selected. The extent to which functional suboptimal binding sites can occur is modeled by a single selection parameter whose exact value, however, remains unknown.

This reasoning leads to a formalism in which DNA motifs bound by a particular TF are represented as a position weight matrix (PWM). As is often, but not always, justified (7, 13), additivity of the binding energy for each base pair is assumed. Provided that the selection model assumptions are satisfied, the discrimination energy associated with the TF-DNA interaction at a given position in the binding site is proportional to the logarithm of the ratio between the frequency in the PWM and the a priori (background) frequency for each nucleotide (83). These pseudo-energies can be represented in terms of a position-specific scoring matrix (PSSM). Virtually all existing computational methods for weight matrix discovery are based on this formalism (3, 51, 79, 84), and the definition of a suitable model for what represents background sequence is a fundamental part of the analysis. The weight matrices summarize the statistical properties of a collection of TF binding sites and therefore represent DNA sequences. Because discrimination energies are inferred up to an unknown scaling factor, a PSSM can only be used to rank sequences by their affinity for the TF.

From a biophysical point of view this formalism is unsatisfying. Weight matrices should represent the properties of the DBD of a TF, not the properties of DNA motifs.

QUANTIFYING TRANSCRIPTION FACTOR–DNA BINDING

Consider a transcription factor (TF) P binding to a DNA sequence S to form the TF-DNA complex PS:

$$P + S \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} \text{PS} \qquad 1.$$

The affinity of the TF for the sequence can be expressed in terms of its equilibrium dissociation constant $K_d(S)$:

$$K_d(S) = \frac{[P][S]}{[PS]} = \frac{k_{\text{off}}}{k_{\text{on}}} = e^{\Delta G/RT}, \qquad 2.$$

which is directly related to ΔG, the Gibbs free energy of binding per mole (where R is the gas constant and T is temperature). The occupancy $N(S)$ of sequence S by transcription factor P can be expressed as the concentration of TF-DNA complex divided by the total concentration of DNA (bound or unbound):

$$N(S) = \frac{[PS]}{[PS]+[S]} = \frac{[P]}{[P]+K_d(S)}. \qquad 3.$$

This equation defines how TF occupancy, with a value between zero and one, depends on both TF concentration and the binding constant.

Position weight matrix (PWM): contains nucleotide frequencies at each position in a collection of DNA sequences

Position-specific scoring matrix (PSSM): derived from a PWM and a background sequence model, contains estimated discrimination energies in unknown units

TF occupancy: the average proportion of time a segment of nucleic acid is bound by a *trans*-factor

Position-specific affinity matrix (PSAM): contains relative affinities directly related to the actual discrimination energies in physical units

Evolutionary arguments should not be invoked when the goal is to model the physical interactions between a TF and the DNA sequence of one specific organism. Moreover, although the strength of TF-DNA interaction varies with local DNA sequence, it should not depend on the choice of a background model representing the global characteristics of noncoding DNA.

An alternative and purely biophysical approach to inferring TF binding specificity from high-throughput genomics data has recently emerged. Continuing a line of thought started by Stormo et al. (85), Clarke and coworkers (35, 54) laid out a statistical-mechanical framework for interpreting TF occupancy measurements in terms of TF-DNA interaction energies. Several algorithms based on the same statistical-mechanical framework directly infer discrimination energies from in vivo or in vitro TF binding data (22, 28, 29, 86). Djordjevic et al. (22) showed that if one assumes that DNA sites are either unoccupied or saturated, the inference of energy parameters reduces to the problem of finding a classifier that distinguishes between bound and unbound sequences, without the need for a background sequence model. The inferred discrimination energies, however, are still only determined up to an arbitrary scaling factor.

Making a different assumption, that TF concentration is below saturation for all sites in the genome, Foat et al. (29) developed an algorithm that determines discrimination energies by fitting a nonlinear model that extends an earlier motif-based method (14). They represented sequence specificity in the form of a position-specific affinity matrix (PSAM). A similar model was used by Tanay (86). In these cases, the scale of the inferred discrimination energies is exactly known, so that quantitative predictions of relative affinity are possible. These works demonstrate the feasibility of a direct biophysical modeling approach that describes TF-DNA interaction not as a digital process, but as an analog process involving a wide range of relative affinities across the genome. Accurate quantification of relative affinity may be particularly important for modeling the temporal response of TF occupancy to an increase in TF activity, or for modeling the complex interplay among multiple TFs, where multiple small energetic contributions together determine the degree to which a DNA region is occupied by a TF complex.

PSSM Versus PSAM

In this section, we discuss in more technical detail how the biophysically motivated PSAM representation of TF sequence specificity relates to the PSSM representation. First, to model how the affinity $K_a(S)$, defined as the inverse of $K_d(S)$, depends on DNA sequence S, consider a reference sequence S_{ref} with a point mutation to base b at position j, resulting in the mutated sequence S_{mut}. Such a mutation will give rise to an additive change $\Delta\Delta G_{jb}$ in the free energy of binding or, equivalently, a multiplicative change w_{jb} in the affinity:

$$\frac{K_a(S_{\text{mut}})}{K_a(S_{\text{ref}})} \equiv w_{jb} = e^{\Delta\Delta G_{jb}/RT}, \qquad 1.$$

where $\Delta\Delta G_{jb} = \Delta G(S_{\text{ref}}) - \Delta G(S_{\text{mut}})$. The collection of w_{jb} values forms a PSAM.

When generalizing to sequences S_{mut} with more than one point mutation, it is usually assumed that the free-energy contributions for each position in the binding site are independent and therefore additive. Equivalently, we can multiply the w_{jb} values for any nucleotide sequence to obtain a predicted relative affinity:

$$\frac{K_a(S_{\text{mut}})}{K_a(S_{\text{ref}})} = \prod_{j=1}^{L_w} w_{j S_{\text{mut}}(j)}. \qquad 2.$$

Here $S_{\text{mut}}(j)$ is the j^{th} base in the mutated sequence, which has length L_w. Note that by definition $w_{j S_{\text{ref}}(j)} = 1$ for all $j = 1, \ldots, L_w$.

Under the evolutionary assumptions made by Berg & Von Hippel (8), an approximate relationship holds between the nucleotide frequencies f in a PWM and the relative affinities w in a PSAM. If the a priori (background) frequency of base b at any position j is given by p_b, then the frequencies in the PWM are given by

$$f_{jb} \propto (w_{jb})^{\lambda} p_b.$$

Here λ is a selection parameter, introduced by Berg & Von Hippel (8), whose exact value is usually unknown. The case when the selection rate is proportional to affinity ($\lambda = 1$) would apply to in vitro selection in the absence of saturation.

Using Structural Information

Structural information from X-ray crystallography is available for a growing number of TFs and TF-DNA complexes in different organisms. This structural information opens up new possibilities for determining TF sequence specificity. For instance, information about the DNA sequences bound by various members of a particular structural class of TF can be combined to build a model that predicts the sequence specificity of a TF from its amino acid sequence alone (13, 45). Alternatively, using the X-ray structure as a template, researchers can use direct molecular modeling of the TF-DNA interface to compute the change in binding free energy when the DNA sequence is mutated (26, 61). Structure-based classification of protein-DNA interaction surfaces can also provide insight into the determinants of binding specificity (77).

PREDICTING GENE EXPRESSION

The rich network of correlations contained in gene expression data over multiple conditions has been analyzed using many different techniques, including clustering (24), principal components analysis (1), bi-clustering (16), multiple regression (32), probabilistic graphical models (30), and information theory (5). In what follows, we limit ourselves to algorithms that are explicitly capable of predicting mRNA expression levels. There is a wide variety of ways in which such models represent gene regulatory logic. We focus on identifying the main conceptual and technical attributes that distinguish the different algorithms available (see **Figure 3**). The earliest and perhaps most popular class of methods relies on the existence of modules to reduce the dimensionality of the expression data to be modeled. Modules are sets of genes that are coexpressed across various experimental conditions, and presumably are coregulated by a common set of TFs. A more recent class of methods takes a fundamentally different approach by using the network of physical interactions between TFs and their targets as a modeling constraint. These methods construct a single model capable of predicting the mRNA expression for any gene in terms of condition-specific regulator activities and gene-specific regulatory network connectivity.

Module-Based Approaches

One of the earliest uses of mRNA expression profiling across multiple conditions was the annotation of genes of unknown function, based on the partitioning of the set of all genes into disjoint clusters with similar expression profiles (24) and the principle of guilt by association. This naturally led to a class of algorithms for modeling gene expression regulation that relied in a fundamental way on the existence of sets of coregulated genes. The earliest studies used Gibbs sampling approaches to look for overrepresented motifs in the promoter regions of coregulated genes (81, 88) without explicitly trying to predict mRNA expression profiles. In Segal et al. (74) and Beer & Tavazoie (6), the predicted mRNA expression profile of a gene is simply the average of all genes in the cluster or module to which it belongs. Both methods start by assigning genes to clusters on the basis of expression only. They subsequently build classifiers that assign genes to clusters on the basis of how likely it is that their promoter region is bound by a particular combination of TFs, and PSSMs are discovered as part of

Reference (algorithm name)	Inferred outputs				Input data			Regulator activity			TF specificity			Method aspects		
	TF activity model	TF specificity	TF susceptibility	Cis-regulatory logic	Sequence	Transcript level	TF occupancy	Regulator transcript level	Module/cluster average	Inferred	Oligonucleotide motif	PSSM	PSAM	Module-centric	Linear model	Regression
Djordjevic et al. 2003 (QPMEME		✓			✓		✓				✓					
Foat et al. 2006 (MatrixREDUCE)		✓			✓		✓						✓			✓
Tanay 2006 (PREGO)		✓			✓		✓						✓			✓
Spellman et al. 1998		✓			✓	✓						✓				
Tavazoie et al. 1999		✓			✓	✓			✓			✓				
Segal, Yelenksy, et al. 2003		✓			✓	✓			✓			✓		✓		
Beer & Tavazoie 2004		✓		✓	✓	✓			✓			✓				
Pe'er et al. 2002 (Minreg)					✓		✓	✓						✓		
Segal, Shapira, et al. 2003		✓			✓	✓		✓				✓		✓		
Middendorf et al. 2004 (GeneClass)					✓	✓		✓			✓					
Kundaje et al. 2006 (Robust GeneClass)					✓	✓	✓	✓				✓				
Chiang et al. 2001		✓			✓	✓			✓		✓					
Bussemaker et al. 2001 (REDUCE)	✓	✓			✓	✓				✓	✓				✓	✓
Keles et al. 2002	✓	✓			✓	✓				✓	✓				✓	✓
Wang et al. 2002	✓	✓			✓	✓				✓	✓				✓	✓
Conlon et al. 2003 (Motif Regressor)	✓	✓			✓	✓				✓		✓			✓	✓
Tanay & Shamir 2004	✓	✓			✓	✓				✓		✓				
Nachman et al. 2004	✓					✓	✓			✓						
Liao et al. 2003 (NCA)	✓		✓			✓	✓			✓					✓	
Tran et al. 2005 (gNCA)	✓		✓			✓	✓			✓					✓	
Nguyen & D'haeseleer 2006 (MED)	✓		✓			✓	✓			✓		✓			✓	
Gao et al. 2004 (MA-Networker)	✓		✓			✓	✓			✓					✓	✓
Bar-Joseph et al. 2003 (GRAM)	✓		✓			✓	✓		✓					✓		
Das et al. 2004 (MARSMotif)	✓	✓		✓	✓	✓				✓	✓	✓				✓
Das et al. 2006	✓	✓		✓	✓	✓				✓		✓				✓
Foat et al. 2005 (MatrixREDUCE)	✓	✓			✓	✓				✓			✓			✓

Figure 3

Classification of several existing algorithms for predicting gene expression on the basis of various attributes. Papers are listed in the order in which their algorithms are described in the text.

the process. On the one hand, the net effect of a particular combination of TFs is implicitly represented by the average mRNA expression profile of each module. On the other hand, the TF binding site combinations that give rise to a given class of expression behavior are represented explicitly. These methods require expression data for a large number of conditions. Furthermore, the expression profile of a significant fraction of the yeast genome cannot be assigned to any module.

A complementary class of module-based methods uses only mRNA expression data as input. Their distinguishing feature is that they use the mRNA expression profile of genes known to encode regulators such as TFs and protein kinases to explain the expression level of the modules. Here again, expression data over a large number of conditions is required. An early embodiment of this approach by Pe'er et al. (65) used an efficient algorithm combined with an objective function based on mutual information to identify an optimal set of key regulators for each gene. In later work, Segal et al. (73) grouped similarly regulated genes into modules and used decision trees to define the predicted expression level of genes in a particular module as a function of the expression level of the regulators controlling the module. These algorithms are capable of predicting the expression of genes in the training set for unobserved conditions given the mRNA levels of the regulatory genes. Promoter sequence and PSSMs are not used to define module membership; therefore the *cis*-regulatory logic that governs the behavior of each module is not explicitly modeled. Nevertheless, one still expects that the promoters of the genes in a given module share certain sequence features, and this can be confirmed by a posteriori sequence analysis.

The two classes of algorithms described above rely fundamentally on the existence of modules of coregulated genes to implicitly represent either (*a*) the condition-specific activity of key regulators or (*b*) the *cis*-regulatory logic used by the transcriptional machinery of the cell to interpret promoter sequences. Middendorf, Kundaje, and colleagues (49, 60) introduced an approach that does not rely on modules and builds a single decision tree model shared by all genes, not just one module. This approach predicts mRNA expression of any gene on the basis of the explicit condition-specific mRNA expression levels of various key regulatory genes and the explicit *cis*-regulatory content of the gene's promoter region. With this approach, researchers can predict the expression level of a gene for which no measurements are available in any condition, knowing only the expression level of the putative key regulators.

Cis-regulatory logic: the rules according to which regulatory signals represented by multiple *trans*-factors binding to the same *cis*-regulatory region of a gene are combined

Transfactome: the combined condition-specific posttranslational regulatory activities of all different *trans*-factors

Modeling TF Activities as Hidden Variables

Fundamentally different from the module-based approach is the use of parametric models that predict the expression level of a gene directly from its promoter sequence. These methods treat the posttranslational activity of each TF as a hidden variable whose value must be estimated from the mRNA expression levels of its target genes. Most expression profiling studies focus on changes in genome-wide mRNA abundance in response to a genetic or a physiological perturbation. Changes in the activity of signaling pathways are communicated to the transcriptional machinery through a distinct layer of control consisting of TFs that bind to DNA. The natural way to model how transcription rates change with cell state is to parameterize TF activity explicitly. We refer to the condition-specific set of TF activities as the transfactome (see **Figure 4**).

In general, the transcription rate of every gene is a highly complex, nonlinear function of the activity of the various TFs that control it. Fortunately, we do not need to know this function if we are only interested in modeling the effect of changes in TF activity on mRNA abundance relative to a reference state. Provided the changes in TF activity are not too large, a first-order, linear model will provide

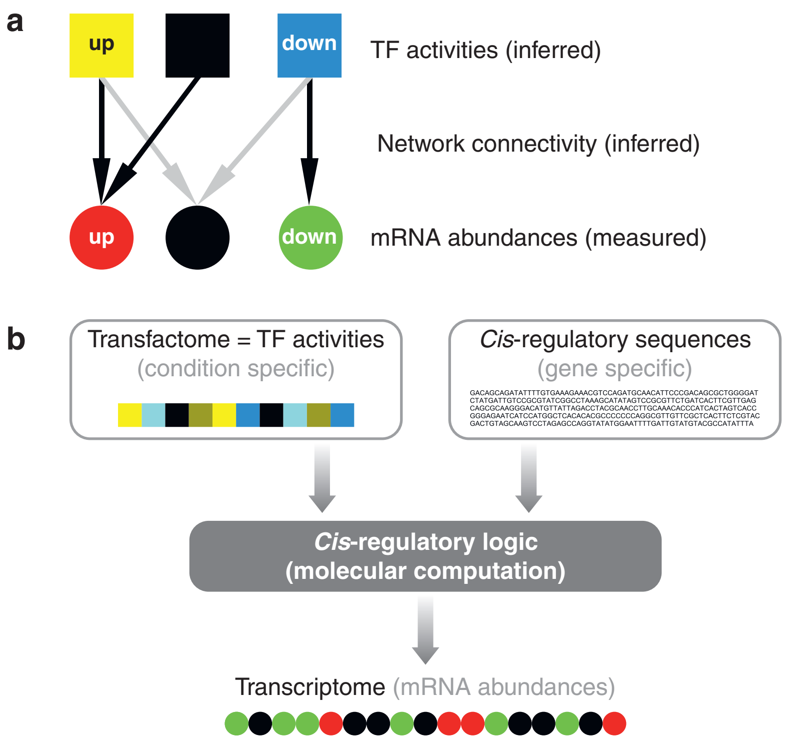
a
up
down
TF activities (inferred)
Network connectivity (inferred)
up
down
mRNA abundances (measured)
b
Transfactome = TF activities
(condition specific)
Cis-regulatory sequences
(gene specific)
GACAGCAGATATTTTGTGAAAGAAACGTCCAGATGCAACATTCCCGACAGCGCTGGGGAT
CTATGATTGTCCGCGTATCGGCCTAAAGCATATAGTCCGCGTTCTGATCACTTCGTTGAG
CAGCGCAAGGGACATGTTATTAGACCTACGCAACCTTGCAAACACCCATCACTAGTCACC
GGGAGAATCATCCATGGCTCACACACGCCCCCCCAGGCGTTGTTCGCTCACTTCTCGTAC
GACTGTAGCAAGTCCTAGAGCCAGGTATATGGAATTTTGATTGTATGTACGCCATATTTA
Cis-regulatory logic
(molecular computation)
Transcriptome (mRNA abundances)

Figure 4

(*a*) Three layers of gene expression control. The degree to which gene-specific mRNA expression levels (*red/green circles*, *bottom row*) respond to condition-specific changes in transcription factor (TF) activity (*yellow/blue squares*, *top row*) is quantified by the regulatory susceptibility associated with each TF-target pair (*arrows*, *middle row*). (*b*) The expression levels of all genes (the transcriptome) depend on the physiological state of the cell through the regulatory activities of all TFs (the transfactome) shared among all genes; each gene is controlled by a different noncoding sequence that determines its susceptibility to changes in TF activity.

a reasonable numerical approximation. Note that this argument does not rely on an assumption of independence among TFs.

The degree to which the expression of a gene is controlled by a TF is determined, at least in part, by the presence of high-affinity binding sites for that TF in the *cis*-regulatory region of the gene. Jensen & Knudsen (44) used nonparametric statistics to detect biases in motif distribution in a set of genes ranked by expression level. Chiang et al. (17) visualized the posttranslational activity of TFs across different microarray experiments by calculating a mean expression profile for all the genes whose regulatory regions contained a specific sequence motif. The use of sequence-based linear regression was introduced by Bussemaker et al. (14), who used forward variable selection to build a linear model for the expression of each gene in terms of the counts of regulatory motifs in its promoter sequence. In the process, Bussemaker et al. (14) determined multivariate linear regression coefficients that estimate the changes in condition-specific posttranslational activity of the TFs that bind to the regulatory motifs. Keles et al. (46) extended this approach by incorporating information about the preferred location of motifs within promoters and using a more elaborate parameter selection method. Wang et al. (93) refined the sequence-based model by multiplying motif counts by expression levels from TF perturbation experiments (e.g., gene deletions) to filter out nonfunctional motif occurrences. Rather than use motifs to represent binding sites, Conlon et al. (18) built their model from a library of PSSMs derived by applying the algorithm MDscan (56) to the regulatory sequences of the genes with the largest changes in expression. Of course, not all algorithms for inferring TF activities as hidden variables use regression. Tanay & Shamir (87) and Nachman et al. (63) developed expectation maximization algorithms that infer posttranslational TF activities and promoter-specific TF affinities from measured mRNA levels.

Omes Law: Linear Response Theory for Genes

A useful analogy exists between gene expression regulation and electricity theory. In Ohm's Law, the susceptibility that relates the current through a resistor to the voltage across the resistor is known as the conductance. In the context of gene expression regulation, TF activity plays the role of the voltage, and the mRNA level that of the current. It is much easier to determine the conductance of a resistor empirically than to compute it from the material properties and geometry of the resistor. Similarly, it is much easier to empirically model a gene's regulatory susceptibility by looking across a large number of experimental conditions than to explicitly predict it from the DNA sequence of its promoter region.

Let A be a matrix of relative mRNA abundances whose rows correspond to genes g and whose columns correspond to conditions c; this would be the usual expression matrix represented by a red-black-green color scheme (24). Each column of A is a transcriptome for a particular condition; each row is the mRNA expression profile of a particular gene. When the activity of one or more TFs changes and the connectivity of the network between TFs and target genes is defined, the response, to linear approximation, is given by the following equation (Omes Law):

$$A_{gc} = \sum_f N_{gf} F_{fc}. \qquad 3.$$

Here N is a network connectivity matrix whose elements represent each gene's susceptibility to change in the activity of each TF; its rows again correspond to genes, but its columns correspond to TFs f. Matrix F contains the TF activity changes; its rows correspond to TFs, and its columns correspond to conditions. Each column of F is a transfactome for a particular condition, each row is the activity profile of a particular TF.

Transcriptome: the concentrations of all different mRNAs

Inferring Regulatory Network Connectivity

In the linear regression methods discussed so far, the predicted affinity of a particular promoter region for a TF serves as a measure of regulatory susceptibility, i.e., the degree to which the expression level of the corresponding gene responds to changes in TF activity. However, as discussed above, TFs do not bind to DNA in isolation. Instead, they do so in complex with a variety of cofactors, each with its own DNA-binding specificity. It is therefore essential to consider the *cis*-regulatory context of each TF binding site, defined by its flanking sequence and possibly more distal elements. Additional information is required to determine to what extent a specific TF binding site contributes to the control of the nearby gene.

A recently developed class of methods infers a matrix of regulatory susceptibilities between genes and TFs by analyzing mRNA expression profiles across a large number of conditions. Following early ideas by Ihmels et al. (41), several researchers (53, 64, 91) have used this approach to estimate gene-specific regulatory susceptibilities. They all begin with a guess about the regulatory connectivities between each TF and all genes in the genome based on experimental TF occupancy data or promoter sequence. Subsequently, mRNA expression data is analyzed to arrive at a

Phylogenetic footprinting: a method for identifying functional *trans*-factor binding sites that looks for conserved noncoding sequences among closely related species

self-consistent pair of matrices F and N representing (*a*) the activity of each TF in each condition and (*b*) the susceptibility of each gene to changes in the activity of each TF, respectively.

For a given network connectivity matrix N, each column of the TF activity matrix F can be estimated by minimizing the mean-squared error between the measured values A and predicted values NF of the change in mRNA expression:

$$F_{\cdot c} = \arg\min \sum_{g} \left(A_{gc} - \sum_{f} N_{gf} F_{fc} \right)^2 . \qquad 4.$$

If motif counts in promoter regions are the elements of matrix N and only a single condition is considered, this procedure is identical to that of Bussemaker et al. (14); ChIP-chip log ratios (31, 53) and PSSM-based affinity scores (18, 64) have also been used to define matrix N.

When expression data for multiple conditions is available, it is possible to re-estimate each row of the susceptibility matrix N on the basis of the inferred TF activity matrix F:

$$N_{g\cdot} = \arg\min \sum_{c} \left(A_{gc} - \sum_{f} N_{gf} F_{fc} \right)^2 . \qquad 5.$$

Gao et al. (31) used this procedure to show that, on average, only 58% of the genes whose promoter region was bound by a TF according to Lee et al. (52) are true regulatory targets. This shows the value of integrating ChIP-chip with mRNA expression data over multiple conditions to define regulatory network connectivity. Bar-Joseph et al. (4) also integrated expression with ChIP-chip data, but pursued the complementary goal of increasing the number of predicted TF targets by using coexpression as additional evidence that a given gene is targeted.

The full bi-linear problem is underdetermined, and iteration between estimating F and N therefore does not lead to a stable, self-consistent solution. However, as was first shown by Tran et al. (91) using ChIP-chip data and later by Nguyen & D'haeseleer (64) using PSSMs and promoter sequence, the problem of simultaneously inferring N and F is rendered well defined by the addition of a weak bias term to the objective function that tries to keep the network topology N either sparse or close to an initial guess.

Using Cross-Species Conservation

Comparative genomics approaches, including both phylogenetic footprinting of orthologous regulatory regions (23) and multispecies techniques that do not rely on linear alignment of DNA sequence (25, 38, 76), have come to dominate the recent literature on *cis*-regulatory analysis. Such approaches are indeed of considerable practical value in weeding out nonfunctional noncoding sequence (57). One could argue, however, that a framework for modeling gene expression regulation based on biophysical principles should not rely on sequence from species other than the one under consideration. Neither the use of evolutionary conservation nor that of linear modeling of regulatory susceptibility across multiple conditions provides an explicit mechanistic explanation for the dependence on *cis*-regulatory context. Neither approach therefore can be used to predict whether a nonfunctional TF binding site can be made functional, or vice versa, by changing the noncoding sequence surrounding the TF binding site. For this, explicit modeling of the complex interplay between multiple TFs and other chromatin-associated proteins binding to the same region is needed.

Complex *Cis*-Regulatory Logic

Pioneering in-depth mutational analysis of complex promoter regions by Arnone & Davidson (2) has helped shed light on the design principles of complex *cis*-regulatory modules (CRMs; not to be confused with the gene sets called modules discussed above). Several groups have computationally predicted CRMs from sequence alone by using sets of functionally related PSSMs as inputs to algorithms designed to detect spatial clusters of TF binding sites (9, 36, 38, 58, 66, 69, 78). A few studies have tried to incorporate spatial and temporal variation of TF activity into their models for complex promoters (43, 71).

Other techniques that can capture aspects of complex regulatory logic include the module-based approach of Beer & Tavazoie (6), which uses a rich *cis*-regulatory grammar that takes into account binding site affinity, position, order, and spacing of TF binding sites. Pilpel et al. (67) identified synergistic motif combinations on the basis of expression coherence. Das et al. (19, 20) systematically explored how the inclusion of terms modeling TF-TF interaction in a sequence-based regression framework can capture aspects of complex regulatory logic in yeast and humans. Motif-based approaches have also been used to identify cofactors of a given TF on the basis of in vivo occupancy data (15, 79).

Theoretical investigations have shown that various types of combinatorial logic can be implemented in terms of the statistical mechanics of interacting proteins and DNA (10, 12). Wang et al. (94) explicitly modeled the regulatory interplay between two yeast TFs binding to overlapping DNA sites. It will also be crucial to address the interplay between TFs and nucleosomes (72). Looking forward, we see biophysical models as among the most natural and promising for modeling complex *cis*-regulatory logic, provided that researchers couple theoretical insight with data-driven estimation of model parameters.

Regulation of mRNA Stability

Most attempts to model mRNA expression have focused on transcription control. Only a handful of studies have taken computational approaches to identify the determinants of mRNA stability regulation. Several conserved sequence motifs were identified downstream of coding regions (47, 95), indicating a likelihood that they are involved in regulating the stability or localization of mRNAs. A few genome-wide studies of mRNA stability have found correspondences between sequence motifs and measured half-lives (37, 75, 96). Gerber et al. (33) measured genome-wide association with mRNA to characterize the sequence specificity of three RNA-binding proteins (RBPs) of the pumilio homology domain (Puf) family in yeast.

RBP: RNA-binding protein

Because they are determined by the balance between transcription and turnover, steady-state mRNA abundances contain implicit information about mRNA decay rates in addition to transcription rates. Wang et al. (93) performed an analysis of nucleotide sequences downstream of genes and identified several oligonucleotide motifs that correlate with changes in steady-state mRNA levels. Sood et al. (80) used a similar approach to analyze tissue-specific regulation by microRNAs. Foat et al. (28) identified PSAMs for two known and a number of as yet unidentified yeast RBPs and demonstrated that their effect on mRNA stability is strongly regulated across hundreds of environmental conditions. They also showed that the transcriptional arrest treatment may change the behavior of mRNA stability regulators, underscoring the advantages of using steady-state expression data.

CONCLUDING REMARKS

Given the many conceptual and technical differences among available methods for modeling mRNA expression, a natural question is: Which ones show the best performance? Such a comparison study was recently performed for motif-finding based on sequence alone (90). Because every method has its own unique way of representing regulatory logic and cell state and uses different types of data as input, it is impossible to compare the models inferred from the data objectively. In our view, the only reasonable way to compare algorithms is through cross-validation: by assessing their ability to predict expression levels that are held out from the dataset on which the algorithms are trained. The comparison should be objective as long as researchers know the scoring function by which the expression values predicted by their method will be compared with the held-out experimental values.

It took physicists many years of painstaking analysis of experimental data and theoretical creativity to arrive at the standard model of particle physics that explains all experimental observations of the interplay between matter and three of the four fundamental forces. They were never able to observe the structure and strength of the interactions described by this model directly, but they could nevertheless estimate its 19 parameters using indirect information provided by scattering events taking place in huge particle accelerators. Molecular biologists are in the early stages of an analogous discovery process, enabled by the genomic revolution. The various current computational modeling strategies may one day converge on a standard model for gene expression regulation based on only the genome sequence and a biophysical description of molecular interactions. Many more than 19 model parameters will need to be determined, but the utility of and understanding generated from such a model will make its construction a worthwhile endeavor.

SUMMARY POINTS

1. A purely biophysical approach to modeling TF-DNA interaction that does not rely on evolutionary assumptions and the use of background sequence models has recently become feasible, owing to the availability of high-throughput TF-DNA interaction data.
2. By modeling TF activities as hidden variables, rather than using the mRNA expression levels of the genes that encode them as a proxy, one can account for the rich posttranslational regulation of TFs.
3. Model-based analysis of mRNA expression profiles over multiple conditions can be used to estimate the regulatory connectivities between TFs and their target genes. This approach provides an alternative to the use of evolutionary conservation to distinguish functional from nonfunctional TF binding sites.
4. Posttranscriptional regulation of transcript stability is an important determinant of steady-state mRNA abundance and can be modeled in close analogy to transcriptional regulation.

FUTURE ISSUES

1. Integration of structural information about TF-DNA interaction with functional genomics data is needed.
2. Investigators will need to model how condition-specific TF activities are modulated by signaling pathways.
3. Explicit modeling of the dependence on *cis*-regulatory sequence context in terms of the complex interactions among multiple TFs and nucleosomes binding to the same DNA region is required.

ACKNOWLEDGMENTS

We thank Gabor Halasz and Ronald Tepper for a critical reading of the manuscript. This work was supported by National Institutes of Health grants R01HG003008, U54CA121852, and R24GM074105.

LITERATURE CITED

1. Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97:10101–6
2. Arnone MI, Davidson EH. 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124:1851–64
3. Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2:28–36
4. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, et al. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21:1337–42
5. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, et al. 2005. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37:382–90
6. Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* 117:185–98
7. Benos PV, Bulyk ML, Stormo GD. 2002. Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Res.* 30:4442–51
8. Berg OG, von Hippel PH. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–50
9. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* 99:757–62
10. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. 2005. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* 15:116–24
11. Blencowe BJ. 2006. Alternative splicing: new insights from global analyses. *Cell* 126:37–47
12. Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA* 100:5136–41
13. Bulyk ML, Johnson PLF, Church GM. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 30:1255–61
14. Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27:167–71
15. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, et al. 2005. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122:33–43
16. Cheng Y, Church GM. 2000. Biclustering of expression data. *Proc. Int. Conf. Intell Syst. Mol. Biol.* 8:93–103
17. Chiang DY, Brown PO, Eisen MB. 2001. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* 17(Suppl. 1):S49–55
18. Conlon EM, Liu XS, Lieb JD, Liu JS. 2003. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* 100:3339–44
19. Das D, Banerjee N, Zhang MQ. 2004. Interacting models of cooperative gene regulation. *Proc. Natl. Acad. Sci. USA* 101:16234–39

20. Das D, Nahl Z, Zhang MQ. 2006. Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.* 2:2006.0029
21. DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–86
22. Djordjevic M, Sengupta AM, Shraiman BI. 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13:2381–90
23. Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* 7:399–406
24. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68
25. Elemento O, Tavazoie S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a nonalignment based approach. *Genome Biol.* 6:R18
26. Endres RG, Wingreen NS. 2006. Weight matrices for protein-DNA binding sites from a single cocrystal structure. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 73:061921
27. Fan J, Yang X, Wang W, Wood WH, Becker KG, et al. 2002. Global analysis of stress-regulated mRNA turnover by using cDNA arrays. *Proc. Natl. Acad. Sci. USA* 99:10611–16
28. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ. 2005. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci. USA* 102:17675–80
29. Foat BC, Morozov AV, Bussemaker HJ. 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22:e141–49
30. Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805
31. Gao F, Foat BC, Bussemaker HJ. 2004. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinform.* 5:31
32. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301:102–5
33. Gerber AP, Herschlag D, Brown PO. 2004. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* 2:E79
34. Gonsalvez GB, Urbinati CR, Long RM. 2005. RNA localization in yeast: moving towards a mechanism. *Biol. Cell* 97:75–86
35. Granek JA, Clarke ND. 2005. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.* 6:R87
36. Gupta M, Liu JS. 2005. De novo *cis*-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 102:7079–84
37. Gutierrez RA, Ewing RM, Cherry JM, Green PJ. 2002. Identification of unstable transcripts in *Arabidopsis* by cDNA microarray analysis: rapid decay is associated with a group of touch- and specific clock-controlled genes. *Proc. Natl. Acad. Sci. USA* 99:11513–18
38. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, et al. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124:47–59
39. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–28
40. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102:109–26
41. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, et al. 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31:370–77

42. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, et al. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409:533–38

43. Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, et al. 2004. Dynamic control of positional information in the early *Drosophila* embryo. *Nature* 430:368–71

44. Jensen LJ, Knudsen S. 2000. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 16:326–33

45. Kaplan T, Friedman N, Margalit H. 2005. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.* 1:e1

46. Keles S, van der Laan M, Eisen MB. 2002. Identification of regulatory elements using a feature selection method. *Bioinformatics* 18:1167–75

47. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–54

48. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. 2005. A high-resolution map of active promoters in the human genome. *Nature* 436:876–80

49. Kundaje A, Middendorf M, Shah M, Wiggins CH, Freund Y, et al. 2006. A classification-based framework for predicting and analyzing gene regulatory response. *BMC Bioinform.* 7(Suppl. 1):S5

50. Lam LT, Pickeral OK, Peng AC, Rosenwald A, Hurt EM, et al. 2001. Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anticancer drug flavopiridol. *Genome Biol.* 2:RESEARCH0041

51. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–14

52. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298:799–804

53. Liao JC, Boscolo R, Yang Y, Tran LM, Sabatti C, et al. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* 100:15522–27

54. Liu X, Clarke ND. 2002. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.* 323:1–8

55. Liu X, Noll DM, Lieb JD, Clarke ND. 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* 15:421–27

56. Liu XS, Brutlag DL, Liu JS. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* 20:835–39

57. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, et al. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinform.* 7:113

58. Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* 99:763–68

59. Mata J, Marguerat S, Bhler J. 2005. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem. Sci.* 30:506–14

60. Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C. 2004. Predicting genetic regulatory response using classification. *Bioinformatics* 20(Suppl. 1):I232–40

61. Morozov AV, Havranek JJ, Baker D, Siggia ED. 2005. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.* 33:5781–98

62. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, et al. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* 36:1331–39

63. Nachman I, Regev A, Friedman N. 2004. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* 20(Suppl. 1):I248–56

64. Nguyen DH, D'haeseleer P. 2006. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol*. 2:2006.0012

65. Pe'er D, Regev A, Tanay A. 2002. Minreg: inferring an active regulator set. *Bioinformatics* 18(Suppl. 1):S258–67

66. Philippakis AA, He FS, Bulyk ML. 2005. Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac. Symp. Biocomput.* pp. 519–30

67. Pilpel Y, Sudarsanam P, Church GM. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*. 29:153–59

68. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122:517–27

69. Rajewsky N, Vergassola M, Gaul U, Siggia ED. 2002. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinform*. 3:30

70. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290:2306–9

71. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, et al. 2004. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* 2:E271

72. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. 2006. A genomic code for nucleosome positioning. *Nature* 442:772–78

73. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. 2003. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34:166–76

74. Segal E, Yelensky R, Koller D. 2003. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 19(Suppl. 1):i273–82

75. Shalgi R, Lapidot M, Shamir R, Pilpel Y. 2005. A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. *Genome Biol.* 6:R86

76. Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* 1:e67

77. Siggers TW, Silkov A, Honig B. 2005. Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol*. 345:1027–45

78. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. 2004. Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinform*. 5:129

79. Smith AD, Sumazin P, Das D, Zhang MQ. 2005. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21(Suppl. 1):i403–12

80. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. 2006. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. USA* 103:2746–51

81. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273–97

82. Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23

83. Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci*. 23:109–13

84. Stormo GD, Hartzell GW. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* 86:1183–87

85. Stormo GD, Schneider TD, Gold L. 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res*. 14:6661–79

86. Tanay A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*. 16:962–72

87. Tanay A, Shamir R. 2004. Multilevel modeling and inference of transcription regulation. *J. Comput. Biol*. 11:357–75

88. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nat. Genet*. 22:281–85

89. Tenenbaum SA, Carson CC, Lager PJ, Keene JD. 2000. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci. USA* 97:14085–90

90. Tompa M, Li N, Bailey TL, Church GM, Moor BD, et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23:137–44

91. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. 2005. gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab. Eng*. 7:128–41

92. van Steensel B, Delrow J, Henikoff S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet*. 27:304–8

93. Wang W, Cherry JM, Botstein D, Li H. 2002. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 99:16893–98

94. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, et al. 2005. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl. Acad. Sci. USA* 102:1998–2003

95. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. *Nature* 434:338–45

96. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, et al. 2003. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* 13:1863–72

97. Yuan G, Liu Y, Dion MF, Slack MD, Wu LF, et al. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309:626–30

98. Zong Q, Schummer M, Hood L, Morris DR. 1999. Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl. Acad. Sci. USA* 96:10632–36

Annual Review of Biophysics and Biomolecular Structure

Volume 35, 2006

Contents

INDEX

ERRATA

An online log of corrections to *Annual Review of Biophysics and Biomolecular Structure* chapters (if any, 1997 to the present) may be found at
http://biophys.annualreviews.org/errata.shtml